Logistic Regression

David Suffolk

5/3/2020

Logistic Regression

Import Data

```
library(caTools)
library(ggplot2)
library(ROCR)
```

```
## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
data <- read.csv("./Project 02/nwCrow_bloodParasites_alaska_smith_2007_2008/nwCrow_sampling_alaska_smit
data_2 <- read.csv("./Project 02/nwCrow_bloodParasites_alaska_smith_2007_2008/nwCrow_bloodParasites_alas
de <- merge(data, data_2, by=0, all=TRUE)
head(de)
```

```
##   Row.names Field.ID      DATE  LOC   LAT    LONG SEX AGE AKD TARSUS WING MASS
## 1         1    75001 3/20/2007 SEWA 60.11 -149.44   1   1   1   55.8  283  448
## 2        10    75010 3/22/2007 KENA 60.55 -151.23   2   1   0   48.6  271  390
## 3       100    75100 3/12/2008 VALD 61.12 -146.35   2   1   0   44.6  264  317
## 4       101    86701 3/12/2008 VALD 61.12 -146.35   2   1   0   47.1  269  343
## 5       102    86702 3/12/2008 VALD 61.12 -146.35   2   2   0   52.2  291  415
## 6       103    86703 3/12/2008 VALD 61.12 -146.35   1   2   0   47.0  266  325
##   Extraction.. LEUC1 LEUC2 HAEM1 HAEM2 PLAS1 PLAS2 Leuc_GenBank_Accession
## 1      NOCR001     0     0     0     0     0     0
## 2      NOCR010     0     0     0     0     0     0
## 3      NOCR100     1     1     0     0     0     0               MG765394
## 4      NOCR101     0     0     0     0     0     0
## 5      NOCR102     1     0     0     0     0     0               MG765394
## 6      NOCR103     1     1     0     0     0     0               MG765394
##   Haem_GenBank_Accession Plas_GenBank_Accession
## 1
## 2
## 3
## 4
## 5
## 6
```

One Hot Encoding

```r
for(unique_value in unique(de$LOC)){


de[paste("LOC", unique_value, sep = ".")] <- ifelse(de$LOC == unique_value, 1, 0)

}
head(de)
```

```
##   Row.names Field.ID      DATE  LOC   LAT    LONG SEX AGE AKD TARSUS WING MASS
## 1         1    75001 3/20/2007 SEWA 60.11 -149.44   1   1   1   55.8  283  448
## 2        10    75010 3/22/2007 KENA 60.55 -151.23   2   1   0   48.6  271  390
## 3       100    75100 3/12/2008 VALD 61.12 -146.35   2   1   0   44.6  264  317
## 4       101    86701 3/12/2008 VALD 61.12 -146.35   2   1   0   47.1  269  343
## 5       102    86702 3/12/2008 VALD 61.12 -146.35   2   2   0   52.2  291  415
## 6       103    86703 3/12/2008 VALD 61.12 -146.35   1   2   0   47.0  266  325
##   Extraction.. LEUC1 LEUC2 HAEM1 HAEM2 PLAS1 PLAS2 Leuc_GenBank_Accession
## 1      NOCR001     0     0     0     0     0     0
## 2      NOCR010     0     0     0     0     0     0
## 3      NOCR100     1     1     0     0     0     0               MG765394
## 4      NOCR101     0     0     0     0     0     0
## 5      NOCR102     1     0     0     0     0     0               MG765394
## 6      NOCR103     1     1     0     0     0     0               MG765394
##   Haem_GenBank_Accession Plas_GenBank_Accession LOC.SEWA LOC.KENA LOC.VALD
## 1                                                      1        0        0
## 2                                                      0        1        0
## 3                                                      0        0        1
## 4                                                      0        0        1
## 5                                                      0        0        1
## 6                                                      0        0        1
##   LOC.HAIN LOC.JUNE LOC.HOME
## 1        0        0        0
## 2        0        0        0
## 3        0        0        0
## 4        0        0        0
## 5        0        0        0
## 6        0        0        0
```

Filter Columns and N/A Values

```r
de <- de[,c(7,8,9,10,11,12,14,16,18,23,24,25,26,27,28)]
de<-de[complete.cases(de),]
head(de)
```

```
##   SEX AGE AKD TARSUS WING MASS LEUC1 HAEM1 PLAS1 LOC.SEWA LOC.KENA LOC.VALD
## 1   1   1   1   55.8  283  448     0     0     0        1        0        0
## 2   2   1   0   48.6  271  390     0     0     0        0        1        0
## 3   2   1   0   44.6  264  317     1     0     0        0        0        1
## 4   2   1   0   47.1  269  343     0     0     0        0        0        1
## 5   2   2   0   52.2  291  415     1     0     0        0        0        1
## 6   1   2   0   47.0  266  325     1     0     0        0        0        1
##   LOC.HAIN LOC.JUNE LOC.HOME
## 1        0        0        0
## 2        0        0        0
## 3        0        0        0
## 4        0        0        0
## 5        0        0        0
## 6        0        0        0
```

AKD

```
mylogit <- glm(AKD ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 + HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + L
summary(mylogit)
```

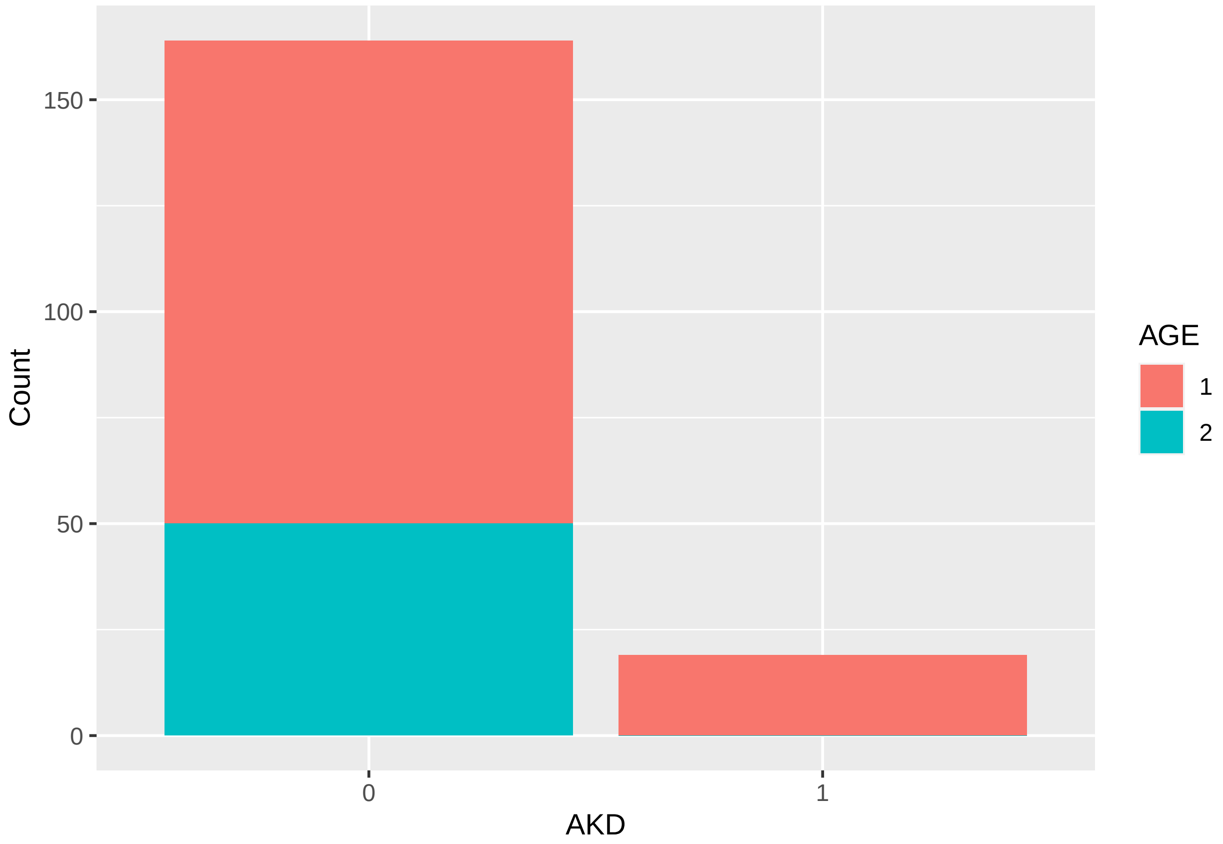
```
##
## Call:
## glm(formula = AKD ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 +
##     HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = de)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.24165  -0.42746  -0.23295  -0.00005   2.32495
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   34.52246 1361.61302   0.025   0.9798
## SEX           -0.55053    0.59753  -0.921   0.3569
## AGE          -18.93752 1361.56638  -0.014   0.9889
## TARSUS        -0.20643    0.18016  -1.146   0.2519
## WING          -0.05237    0.04529  -1.156   0.2475
## MASS           0.01619    0.01366   1.185   0.2358
## LEUC1         -0.24943    0.63514  -0.393   0.6945
## HAEM1          1.18295    0.82572   1.433   0.1520
## PLAS1          0.54607    0.88109   0.620   0.5354
## LOC.SEWA       2.76103    1.24752   2.213   0.0269 *
## LOC.KENA       3.00526    1.28620   2.337   0.0195 *
## LOC.VALD       1.48368    1.24096   1.196   0.2319
## LOC.HAIN      -0.09574    1.56040  -0.061   0.9511
## LOC.JUNE       1.20136    1.30677   0.919   0.3579
## LOC.HOME            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 122.027  on 182  degrees of freedom
## Residual deviance:  89.244  on 169  degrees of freedom
## AIC: 117.24
##
## Number of Fisher Scoring iterations: 18
```

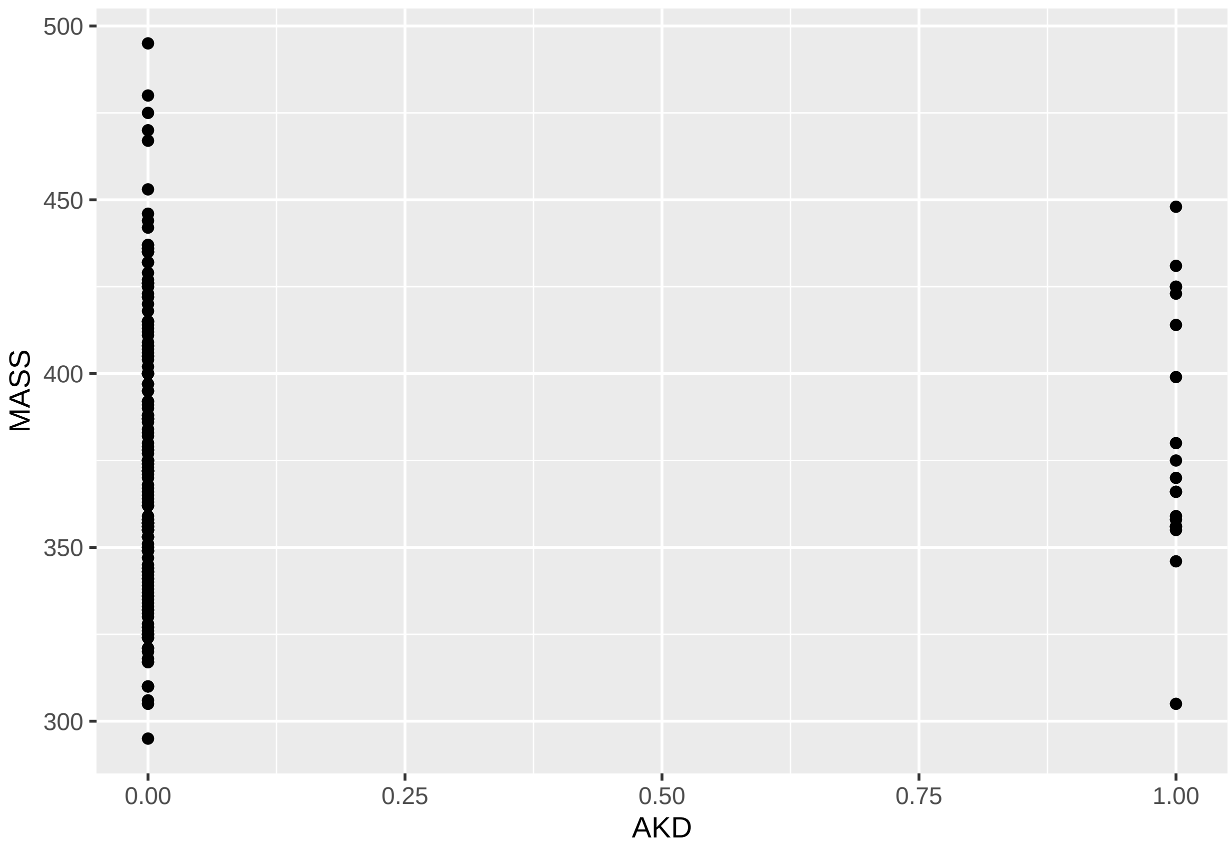
```
confint.default(mylogit)
```

```
##                     2.5 %       97.5 %
## (Intercept) -2634.1900120 2.703235e+03
## SEX            -1.7216634 6.205950e-01
## AGE         -2687.5585861 2.649684e+03
## TARSUS         -0.5595384 1.466764e-01
## WING           -0.1411362 3.639016e-02
## MASS           -0.0105768 4.295700e-02
## LEUC1          -1.4942911 9.954247e-01
## HAEM1          -0.4354375 2.801334e+00
## PLAS1          -1.1808293 2.272975e+00
## LOC.SEWA        0.3159412 5.206111e+00
## LOC.KENA        0.4843588 5.526163e+00
## LOC.VALD       -0.9485564 3.915921e+00
## LOC.HAIN       -3.1540757 2.962598e+00
## LOC.JUNE       -1.3598649 3.762584e+00
## LOC.HOME               NA           NA
```

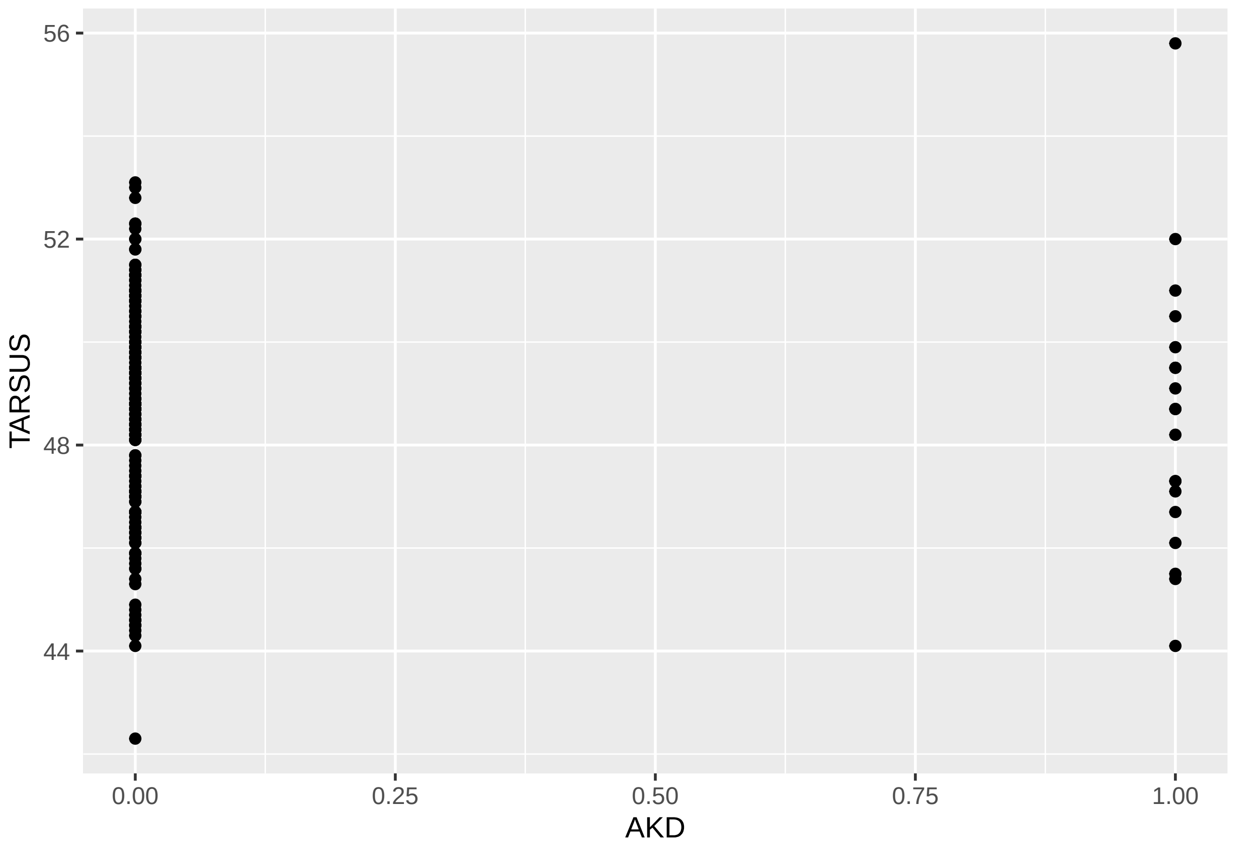
```
dat <- data.frame(table(de$AKD, de$AGE))
names(dat) <- c("AKD","AGE","Count")
ggplot(data=dat, aes(x=AKD, y=Count, fill=AGE)) + geom_bar(stat="identity")
```



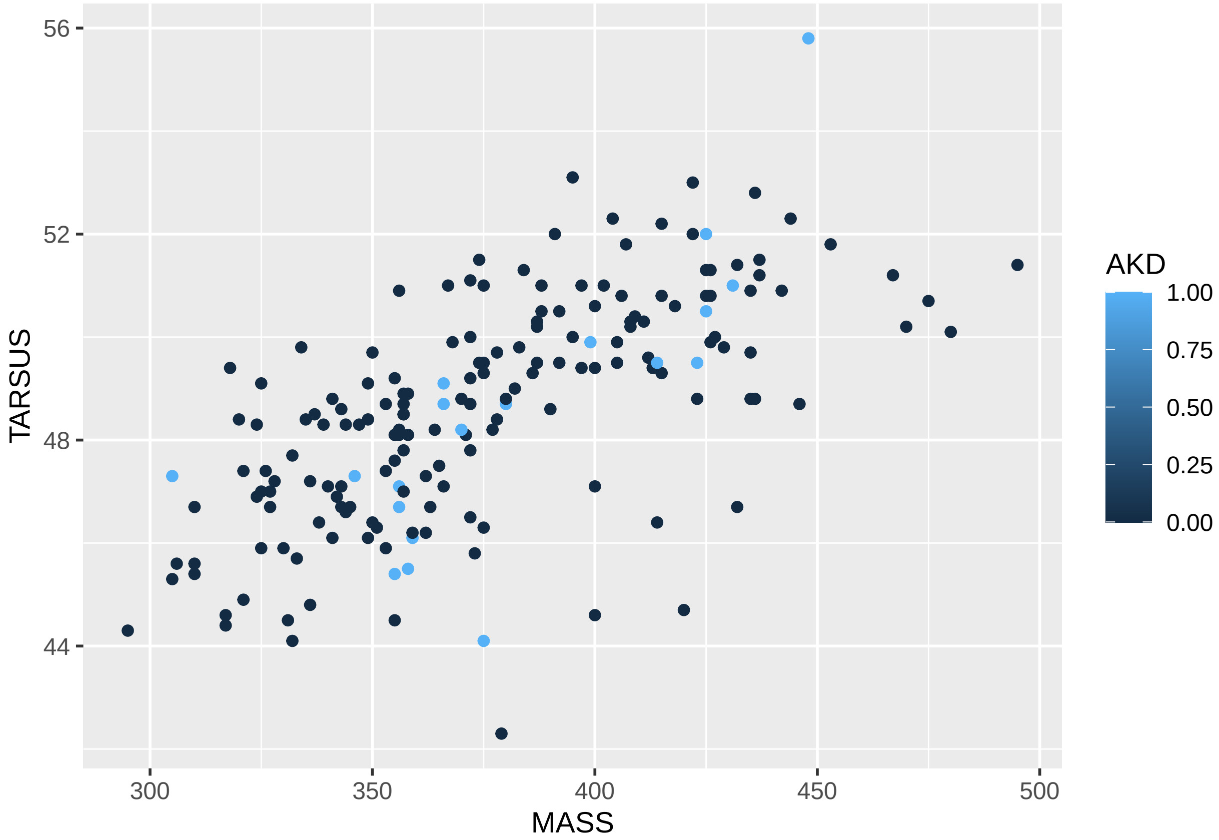
```r
dat <- data.frame(table(de$AKD, de$MASS))
names(dat) <- c("AKD","MASS","Count")
#ggplot(data=dat, aes(x=AKD, y=Count, fill=MASS)) + geom_bar(stat="identity")

ggplot(de, aes(x = AKD, y = MASS)) +
    geom_point()
```



```r
ggplot(de, aes(x = AKD, y = TARSUS)) +
    geom_point()
```



```r
ggplot(de, aes(x = MASS, y = TARSUS, color=AKD)) +
    geom_point()
```



```r
set.seed(88)
split <- sample.split(de$AKD, SplitRatio = 0.75)
dresstrain <- subset(de, split == TRUE)
dresstest <- subset(de, split == FALSE)
model <- glm (AKD ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 + HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + LO
summary(model)
```

```
##
## Call:
## glm(formula = AKD ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 +
##     HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = dresstrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.38244  -0.38308  -0.19102  -0.00006   2.35213
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    12.94147 1549.85315   0.008   0.9933
## SEX            -0.69146    0.89306  -0.774   0.4388
## AGE           -18.20809 1549.78767  -0.012   0.9906
## TARSUS         -0.28295    0.20812  -1.360   0.1740
## WING            0.03222    0.06149   0.524   0.6003
## MASS            0.02012    0.01851   1.087   0.2770
## LEUC1         -0.70013    0.81267  -0.862   0.3889
## HAEM1          1.20972    1.01178   1.196   0.2318
## PLAS1          0.86431    1.04429   0.828   0.4079
## LOC.SEWA       3.04810    1.38552   2.200   0.0278 *
## LOC.KENA       2.60819    1.49195   1.748   0.0804 .
## LOC.VALD       1.32461    1.42093   0.932   0.3512
## LOC.HAIN       0.33166    1.64356   0.202   0.8401
## LOC.JUNE       0.03798    1.56011   0.024   0.9806
## LOC.HOME            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.384  on 136  degrees of freedom
## Residual deviance: 59.804  on 123  degrees of freedom
## AIC: 87.804
##
## Number of Fisher Scoring iterations: 18
```

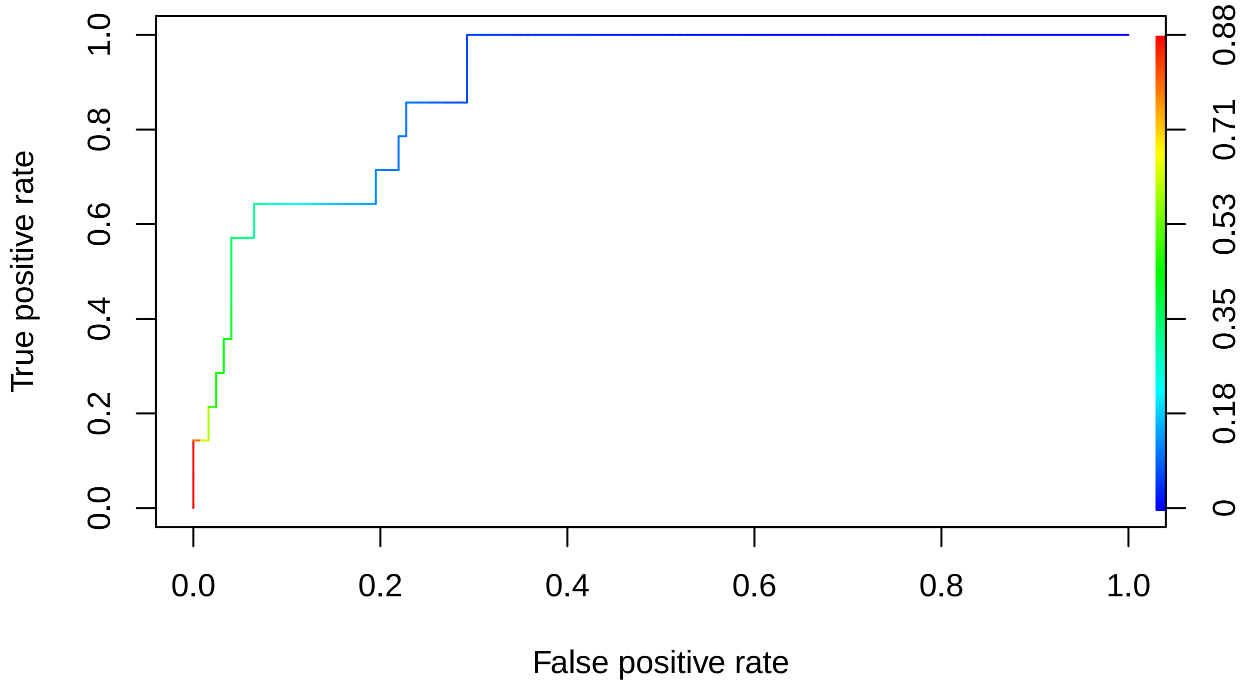
```
predict <- predict(model, type = 'response')
tab2 <- table(dresstrain$AKD, predict > 0.5)
tab2
```

```
##
##     FALSE TRUE
##   0   121    2
##   1    12    2
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 89.78102
```

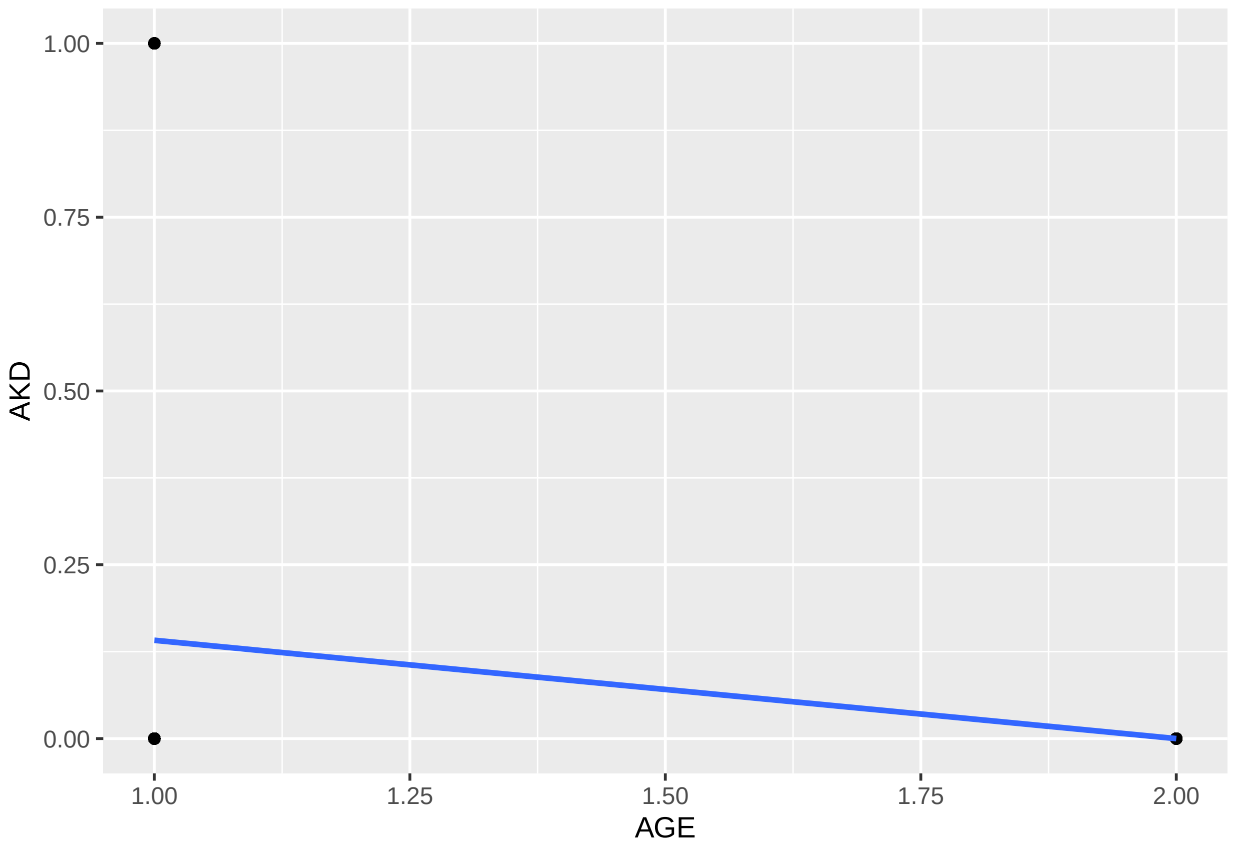
```
#ROCR Curve
ROCRpred <- prediction(predict, dresstrain$AKD)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```



```
#plot glm
ggplot(dresstrain, aes(x=AGE, y=AKD)) + geom_point() +
stat_smooth(method="glm", family="binomial", se=FALSE)
```

```
## Warning: Ignoring unknown parameters: family
```

```
## `geom_smooth()` using formula 'y ~ x'
```



LEUC1

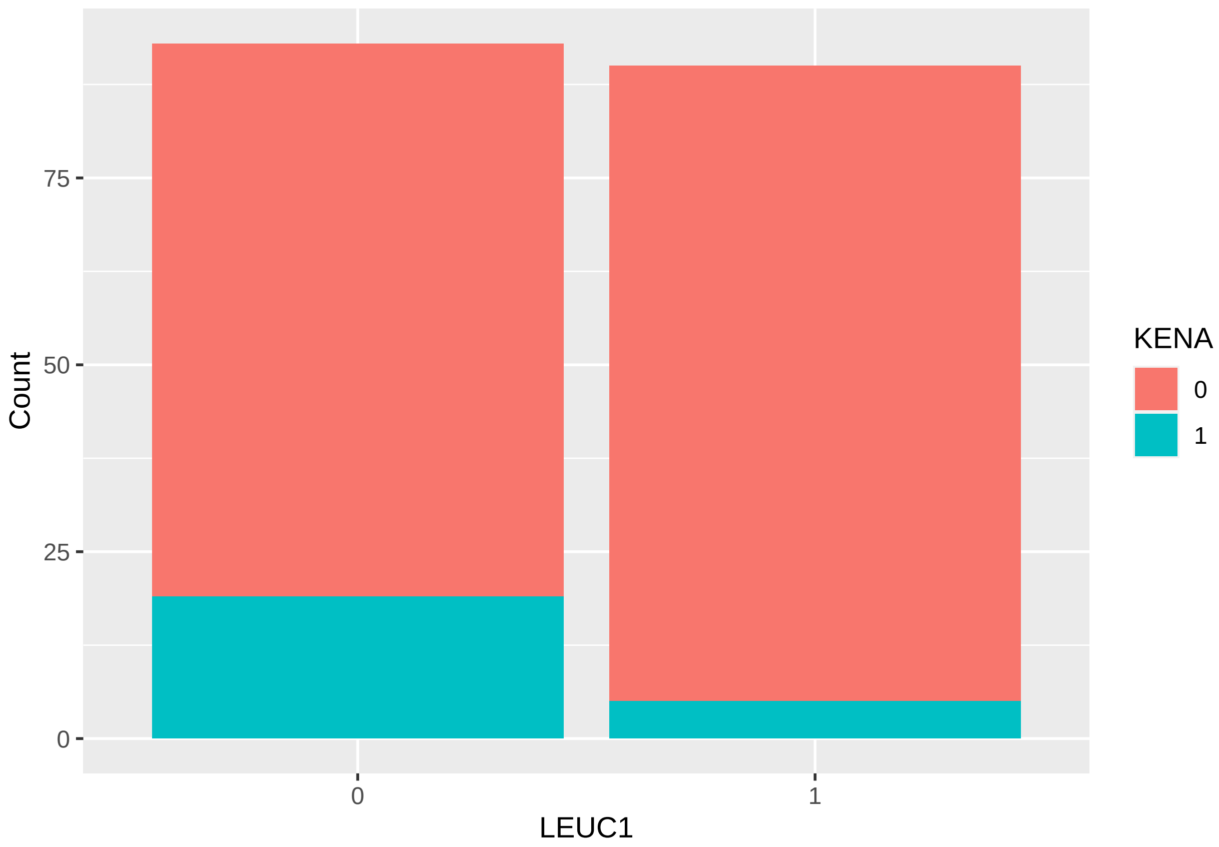
```
mylogit <- glm(LEUC1 ~ SEX + AGE + TARSUS + WING + MASS + AKD + HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + L
summary(mylogit)
```

```
##
## Call:
## glm(formula = LEUC1 ~ SEX + AGE + TARSUS + WING + MASS + AKD +
##     HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = de)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9722  -1.0370  -0.4954   1.0100   1.9081
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.901758   6.026071  -0.979  0.32740
## SEX         -0.634848   0.344574  -1.842  0.06541 .
## AGE          0.169600   0.417676   0.406  0.68470
## TARSUS       0.041890   0.106828   0.392  0.69497
## WING         0.030357   0.025176   1.206  0.22791
## MASS        -0.007899   0.006329  -1.248  0.21202
## AKD         -0.271437   0.572274  -0.474  0.63528
## HAEM1        0.591862   0.432145   1.370  0.17081
## PLAS1       -0.278315   0.582255  -0.478  0.63265
## LOC.SEWA    -1.053987   0.652302  -1.616  0.10614
## LOC.KENA    -2.119886   0.701517  -3.022  0.00251 **
## LOC.VALD    -0.139498   0.562028  -0.248  0.80398
## LOC.HAIN    -1.382313   0.622126  -2.222  0.02629 *
## LOC.JUNE    -1.523207   0.618522  -2.463  0.01379 *
## LOC.HOME           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 253.64  on 182  degrees of freedom
## Residual deviance: 224.69  on 169  degrees of freedom
## AIC: 252.69
##
## Number of Fisher Scoring iterations: 4
```

```
confint.default(mylogit)
```

```
##                    2.5 %       97.5 %
## (Intercept) -17.71263917  5.909124006
## SEX          -1.31020091  0.040505299
## AGE          -0.64902989  0.988229952
## TARSUS       -0.16748996  0.251269057
## WING         -0.01898793  0.079701079
## MASS         -0.02030441  0.004506129
## AKD          -1.39307227  0.850198909
## HAEM1        -0.25512635  1.438850338
## PLAS1        -1.41951384  0.862883216
## LOC.SEWA     -2.33247532  0.224501978
## LOC.KENA     -3.49483430 -0.744937835
## LOC.VALD     -1.24105373  0.962057262
## LOC.HAIN     -2.60165698 -0.162969019
## LOC.JUNE     -2.73548874 -0.310925938
## LOC.HOME              NA           NA
```

```
dat <- data.frame(table(de$LEUC1, de$LOC.KENA))
names(dat) <- c("LEUC1","KENA","Count")
ggplot(data=dat, aes(x=LEUC1, y=Count, fill=KENA)) + geom_bar(stat="identity")
```



```
set.seed(88)
split <- sample.split(de$LEUC1, SplitRatio = 0.75)
dresstrain <- subset(de, split == TRUE)
dresstest <- subset(de, split == FALSE)
model <- glm (LEUC1 ~ SEX + AGE + TARSUS + WING + MASS + AKD + HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + LO
summary(model)
```

```
##
## Call:
## glm(formula = LEUC1 ~ SEX + AGE + TARSUS + WING + MASS + AKD +
##     HAEM1 + PLAS1 + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = dresstrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0594  -0.9543  -0.3221   0.9395   1.9845
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -17.501779   7.846213  -2.231  0.02571 *
## SEX          -0.867256   0.416394  -2.083  0.03727 *
## AGE           0.844796   0.550956   1.533  0.12520
## TARSUS        0.134716   0.129435   1.041  0.29797
## WING          0.068958   0.030971   2.227  0.02598 *
## MASS         -0.018549   0.008013  -2.315  0.02061 *
## AKD           0.336550   0.665383   0.506  0.61300
## HAEM1        -0.236943   0.534565  -0.443  0.65759
## PLAS1        -0.240777   0.673583  -0.357  0.72075
## LOC.SEWA     -0.998717   0.758362  -1.317  0.18786
## LOC.KENA     -2.600282   0.905352  -2.872  0.00408 **
## LOC.VALD      0.151831   0.687561   0.221  0.82523
## LOC.HAIN     -0.903689   0.742812  -1.217  0.22376
## LOC.JUNE     -1.406570   0.732075  -1.921  0.05469 .
## LOC.HOME            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 191.28  on 137  degrees of freedom
## Residual deviance: 157.56  on 124  degrees of freedom
## AIC: 185.56
##
## Number of Fisher Scoring iterations: 4
```

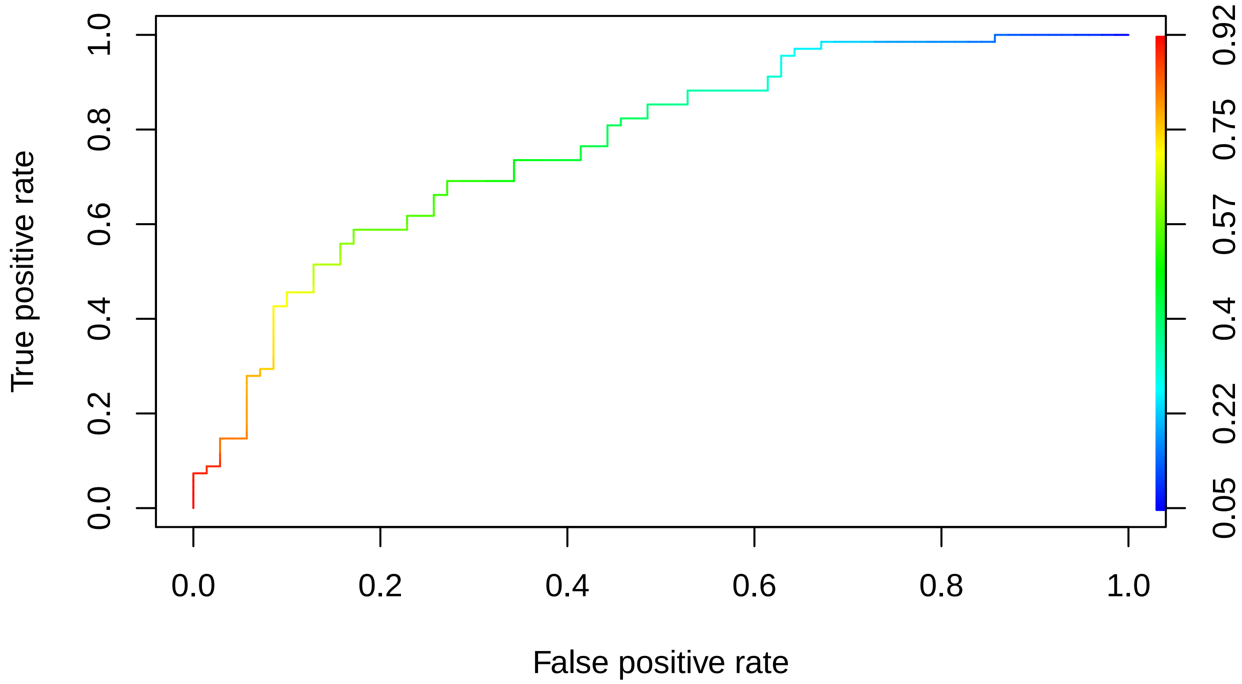
```
predict <- predict(model, type = 'response')
tab2 <- table(dresstrain$LEUC1, predict > 0.5)
tab2
```

```
##
##     FALSE TRUE
##   0    49   21
##   1    21   47
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 69.56522
```

```
#ROCR Curve
library(ROCR)
ROCRpred <- prediction(predict, dresstrain$LEUC1)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```



HAEM1

```
mylogit <- glm(HAEM1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 + AKD + PLAS1 + LOC.SEWA + LOC.KENA + L
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

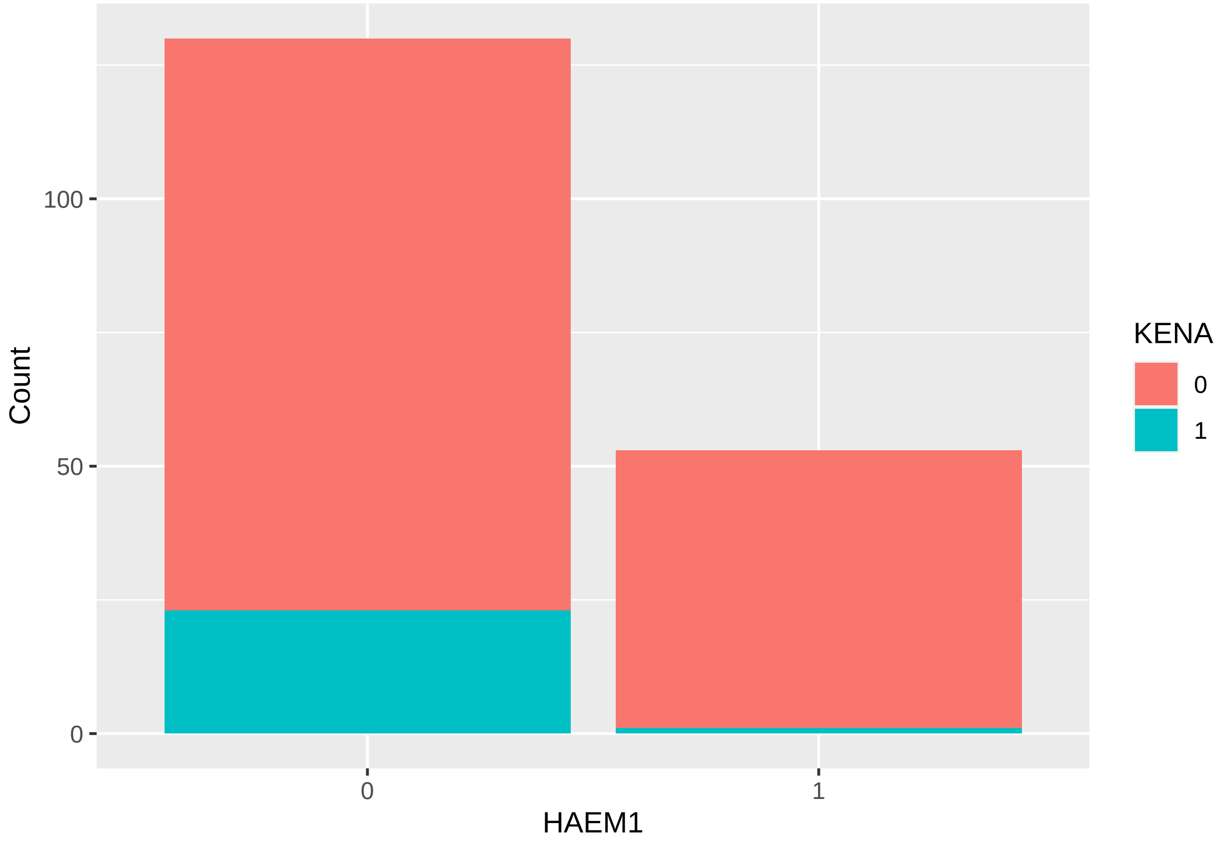
```
summary(mylogit)
```

```
##
## Call:
## glm(formula = HAEM1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 +
##     AKD + PLAS1 + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = de)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9600  -0.6418  -0.0001   0.7594   2.1300
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.328e+00  8.279e+00  -1.127   0.2599
## SEX         -1.112e-01  4.529e-01  -0.246   0.8060
## AGE          1.436e+00  5.857e-01   2.451   0.0142 *
## TARSUS       1.433e-01  1.406e-01   1.019   0.3081
## WING         1.077e-02  3.236e-02   0.333   0.7393
## MASS        -1.044e-02  7.925e-03  -1.318   0.1876
## LEUC1        6.784e-01  4.491e-01   1.511   0.1309
## AKD          6.331e-01  7.522e-01   0.842   0.3999
## PLAS1       -1.807e+01  2.221e+03  -0.008   0.9935
## LOC.SEWA     1.103e+00  8.204e-01   1.344   0.1788
## LOC.KENA    -2.190e+00  1.253e+00  -1.747   0.0806 .
## LOC.VALD    -1.845e+01  1.632e+03  -0.011   0.9910
## LOC.HAIN     1.785e+00  7.233e-01   2.468   0.0136 *
## LOC.JUNE     1.287e+00  7.463e-01   1.724   0.0847 .
## LOC.HOME            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 220.26  on 182  degrees of freedom
## Residual deviance: 137.23  on 169  degrees of freedom
## AIC: 165.23
##
## Number of Fisher Scoring iterations: 18
```

```
confint.default(mylogit)
```

```
##                     2.5 %       97.5 %
## (Intercept) -2.555382e+01 6.897946e+00
## SEX         -9.989220e-01 7.764953e-01
## AGE          2.875900e-01 2.583641e+00
## TARSUS      -1.322820e-01 4.189151e-01
## WING        -5.266542e-02 7.420183e-02
## MASS        -2.597683e-02 5.088832e-03
## LEUC1       -2.017605e-01 1.558561e+00
## AKD         -8.411244e-01 2.107398e+00
## PLAS1       -4.370732e+03 4.334582e+03
## LOC.SEWA    -5.049756e-01 2.711085e+00
## LOC.KENA    -4.645821e+00 2.663481e-01
## LOC.VALD    -3.216350e+03 3.179455e+03
## LOC.HAIN     3.676849e-01 3.202834e+00
## LOC.JUNE    -1.759188e-01 2.749479e+00
## LOC.HOME               NA           NA
```

```
dat <- data.frame(table(de$HAEM1, de$LOC.KENA))
names(dat) <- c("HAEM1","KENA","Count")
ggplot(data=dat, aes(x=HAEM1, y=Count, fill=KENA)) + geom_bar(stat="identity")
```



```
set.seed(88)
split <- sample.split(de$HAEM1, SplitRatio = 0.75)
dresstrain <- subset(de, split == TRUE)
dresstest <- subset(de, split == FALSE)
model <- glm (HAEM1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 + AKD + PLAS1 + LOC.SEWA + LOC.KENA + LOC
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model)
```

```
##
## Call:
## glm(formula = HAEM1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 +
##     AKD + PLAS1 + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = dresstrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.04529  -0.61211  -0.00008   0.56463   2.03534
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -14.04400   10.35263  -1.357  0.17492
## SEX          -0.26417    0.54237  -0.487  0.62621
## AGE            2.11533    0.74242   2.849  0.00438 **
## TARSUS         0.03114    0.18869   0.165  0.86890
## WING           0.05333    0.04095   1.302  0.19287
## MASS          -0.01585    0.01008  -1.572  0.11587
## LEUC1          0.86884    0.54502   1.594  0.11091
## AKD            1.30502    0.93512   1.396  0.16285
## PLAS1        -18.55806 2271.09884  -0.008  0.99348
## LOC.SEWA       0.88801    0.97061   0.915  0.36024
## LOC.KENA      -2.40842    1.35030  -1.784  0.07449 .
## LOC.VALD     -18.75502 1816.09319  -0.010  0.99176
## LOC.HAIN       1.99819    0.84955   2.352  0.01867 *
## LOC.JUNE       0.23395    0.87129   0.269  0.78830
## LOC.HOME            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.158  on 137  degrees of freedom
## Residual deviance:  96.705  on 124  degrees of freedom
## AIC: 124.7
##
## Number of Fisher Scoring iterations: 18
```

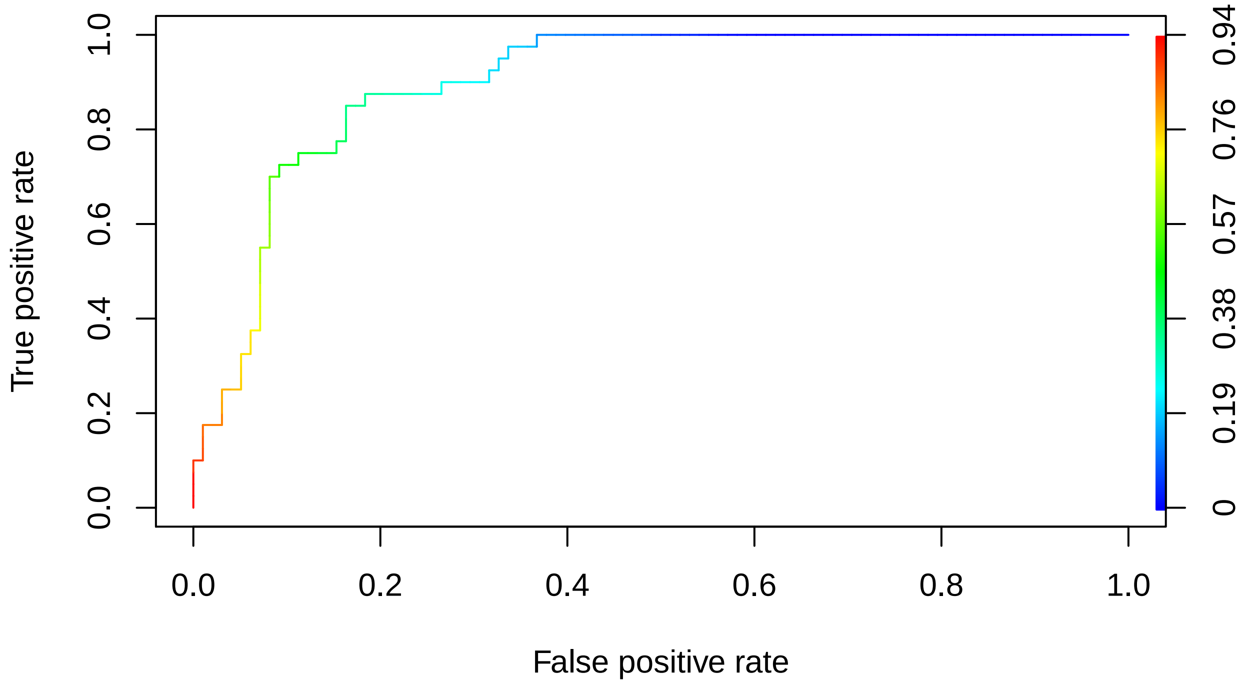
```
predict <- predict(model, type = 'response')
tab2 <- table(dresstrain$HAEM1, predict > 0.5)
tab2
```

```
##
##     FALSE TRUE
##   0    90    8
##   1    12   28
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 85.50725
```

```
#ROCR Curve
ROCRpred <- prediction(predict, dresstrain$HAEM1)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```



PLAS1

```
mylogit <- glm(PLAS1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 + HAEM1 + AKD + LOC.SEWA + LOC.KENA + L
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

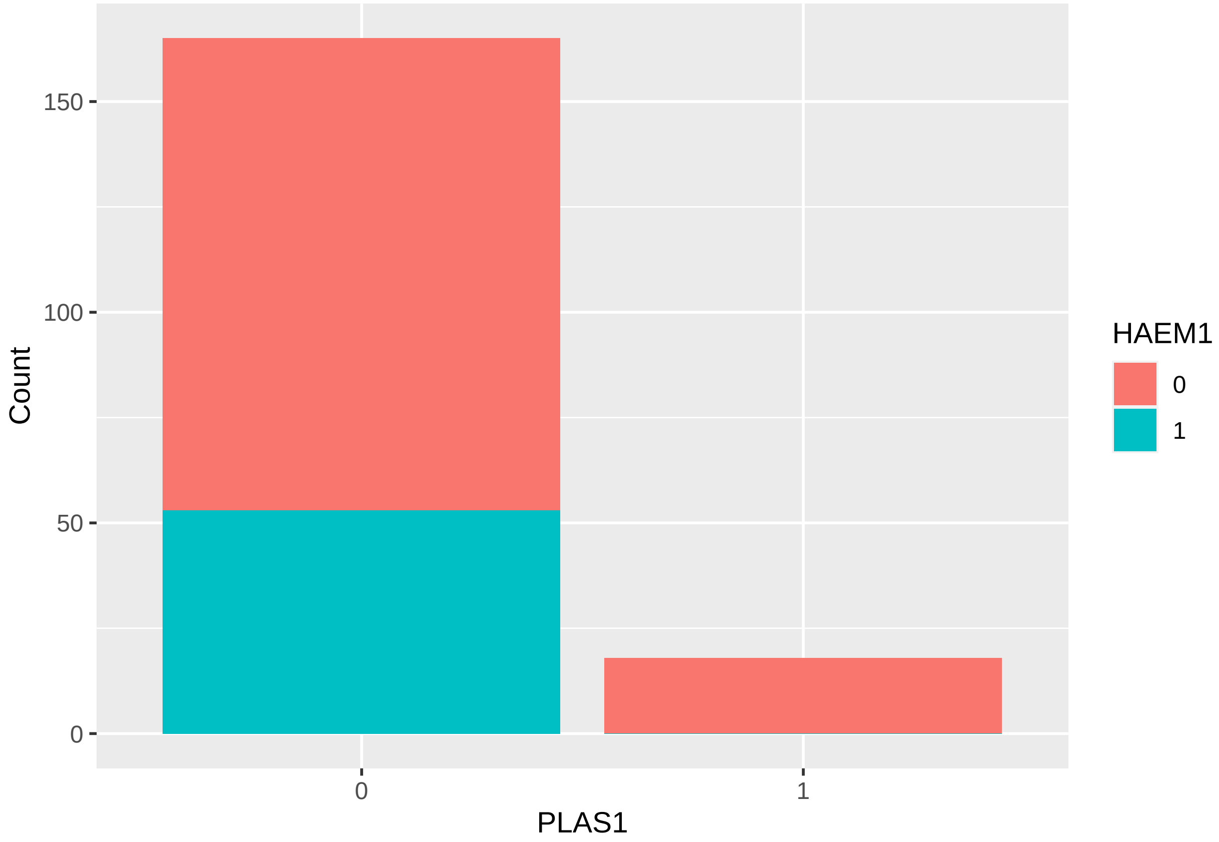
```
summary(mylogit)
```

```
##
## Call:
## glm(formula = PLAS1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 +
##     HAEM1 + AKD + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = de)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.27402  -0.48568  -0.25260  -0.00003   2.27954
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -18.17041   11.26057  -1.614   0.1066
## SEX           -0.02530    0.59058  -0.043   0.9658
## AGE           -0.02371    0.89907  -0.026   0.9790
## TARSUS        -0.14254    0.18544  -0.769   0.4421  
## WING           0.12195    0.05138   2.374   0.0176 *
## MASS          -0.02663    0.01304  -2.042   0.0411 *
## LEUC1         -0.21766    0.61422  -0.354   0.7231  
## HAEM1        -18.77793 2064.99936  -0.009   0.9927  
## AKD            0.95999    0.87938   1.092   0.2750  
## LOC.SEWA       0.52272    1.02112   0.512   0.6087  
## LOC.KENA      -1.52239    1.03462  -1.471   0.1412  
## LOC.VALD      -0.90864    0.73873  -1.230   0.2187  
## LOC.HAIN     -17.75856 2491.36553  -0.007   0.9943  
## LOC.JUNE      -0.52983    0.93841  -0.565   0.5723  
## LOC.HOME            NA         NA      NA       NA  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 117.66  on 182  degrees of freedom
## Residual deviance:  87.48  on 169  degrees of freedom
## AIC: 115.48
##
## Number of Fisher Scoring iterations: 19
```

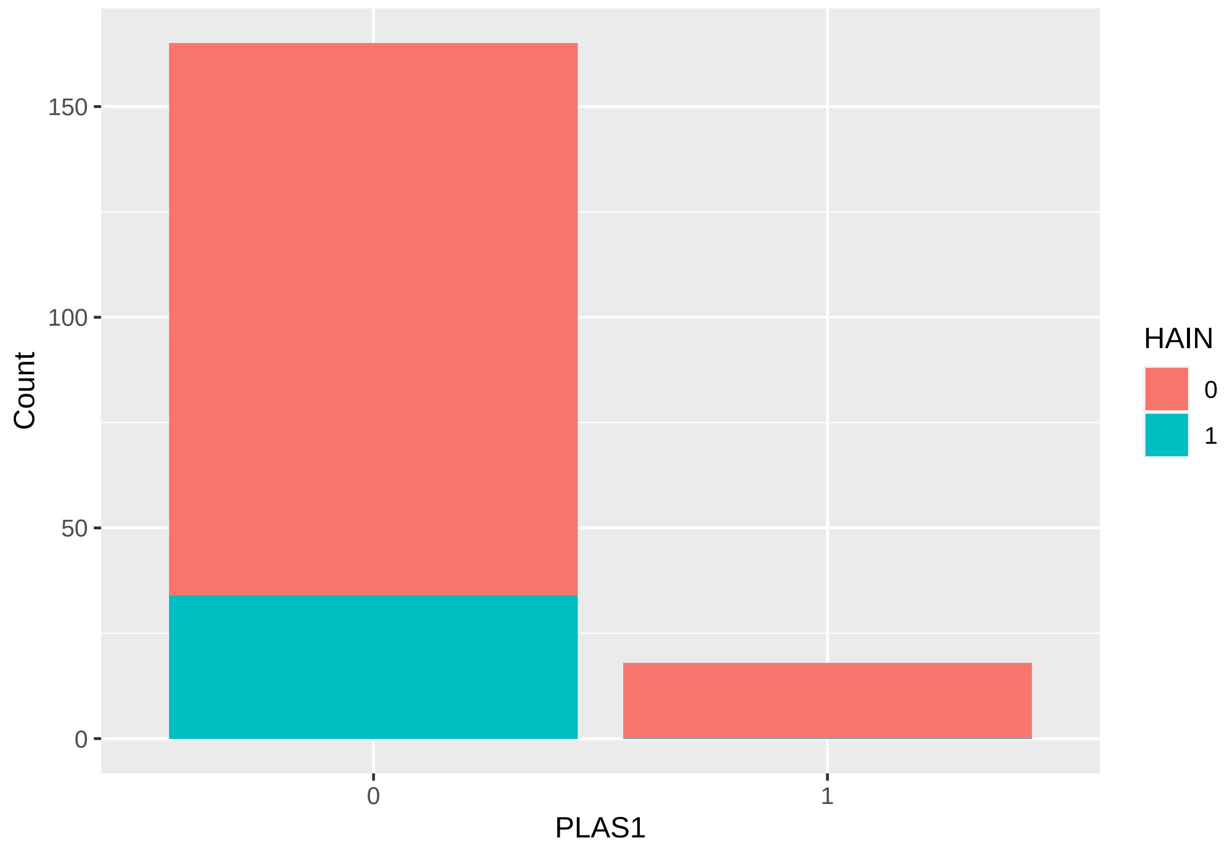
```
confint.default(mylogit)
```

```
##                     2.5 %        97.5 %
## (Intercept) -4.024072e+01  3.899895e+00
## SEX         -1.182815e+00  1.132212e+00
## AGE         -1.785850e+00  1.738433e+00
## TARSUS      -5.059899e-01  2.209062e-01
## WING         2.125323e-02  2.226448e-01
## MASS        -5.219369e-02 -1.072603e-03
## LEUC1       -1.421506e+00  9.861846e-01
## HAEM1       -4.066102e+03  4.028546e+03
## AKD         -7.635654e-01  2.683554e+00
## LOC.SEWA    -1.478637e+00  2.524080e+00
## LOC.KENA    -3.550206e+00  5.054281e-01
## LOC.VALD    -2.356527e+00  5.392387e-01
## LOC.HAIN    -4.900745e+03  4.865228e+03
## LOC.JUNE    -2.369083e+00  1.309431e+00
## LOC.HOME               NA            NA
```

```
dat <- data.frame(table(de$PLAS1, de$HAEM1))
names(dat) <- c("PLAS1","HAEM1","Count")
ggplot(data=dat, aes(x=PLAS1, y=Count, fill=HAEM1)) + geom_bar(stat="identity")
```



```
dat <- data.frame(table(de$PLAS1, de$LOC.HAIN))
names(dat) <- c("PLAS1","HAIN","Count")
ggplot(data=dat, aes(x=PLAS1, y=Count, fill=HAIN)) + geom_bar(stat="identity")
```



```
set.seed(88)
split <- sample.split(de$PLAS1, SplitRatio = 0.75)
dresstrain <- subset(de, split == TRUE)
dresstest <- subset(de, split == FALSE)
model <- glm (PLAS1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 + HAEM1 + AKD + LOC.SEWA + LOC.KENA + LO
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model)
```

```
##
## Call:
## glm(formula = PLAS1 ~ SEX + AGE + TARSUS + WING + MASS + LEUC1 +
##     HAEM1 + AKD + LOC.SEWA + LOC.KENA + LOC.VALD + LOC.HAIN +
##     LOC.JUNE + LOC.HOME, family = "binomial", data = dresstrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.41210  -0.43222  -0.16854  -0.00002   2.49957
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -20.43406   15.22318  -1.342   0.1795
## SEX           -0.69492    0.75744  -0.917   0.3589
## AGE            1.04784    1.08626   0.965   0.3347
## TARSUS        -0.08533    0.22805  -0.374   0.7083
## WING           0.14023    0.06533   2.147   0.0318 *
## MASS          -0.04157    0.01816  -2.290   0.0220 *
## LEUC1         -0.04961    0.82272  -0.060   0.9519
## HAEM1        -19.43245 2253.67577  -0.009   0.9931
## AKD            1.82711    1.03636   1.763   0.0779 .
## LOC.SEWA      -0.35821    1.31096  -0.273   0.7847
## LOC.KENA      -2.90932    1.42365  -2.044   0.0410 *
## LOC.VALD      -1.71265    0.91620  -1.869   0.0616 .
## LOC.HAIN     -18.48739 2732.63500  -0.007   0.9946
## LOC.JUNE      -0.54674    1.03178  -0.530   0.5962
## LOC.HOME            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.599  on 137  degrees of freedom
## Residual deviance: 60.707  on 124  degrees of freedom
## AIC: 88.707
##
## Number of Fisher Scoring iterations: 19
```

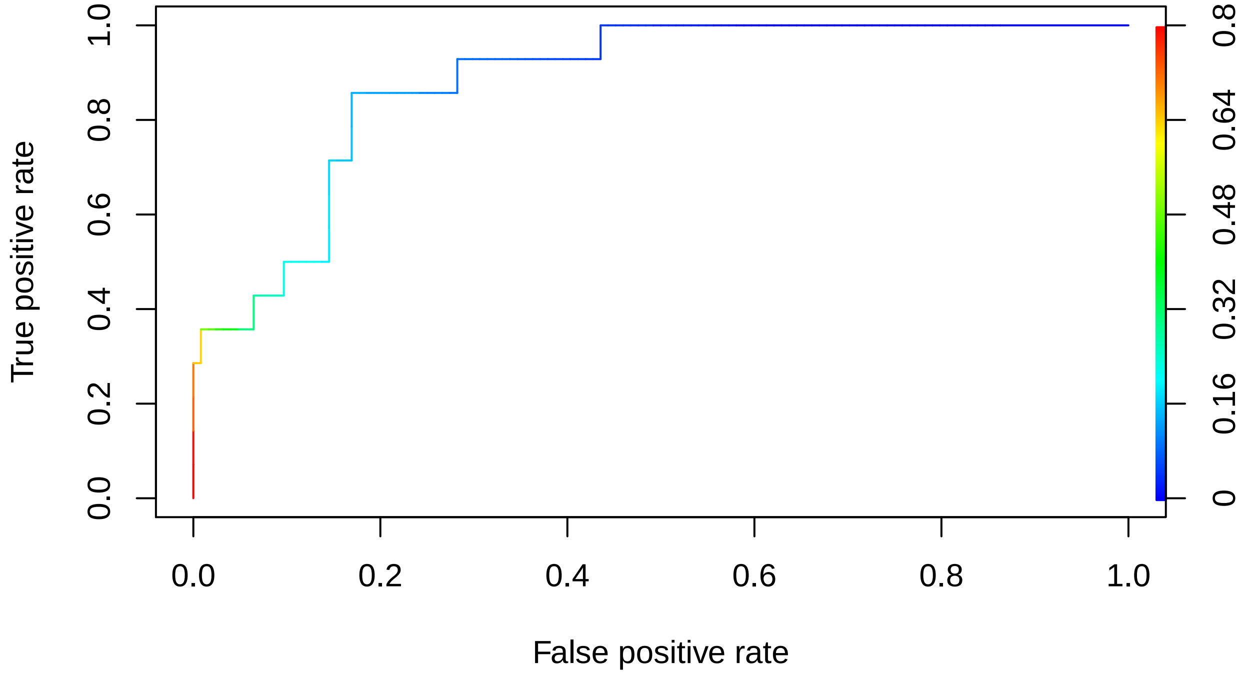
```
predict <- predict(model, type = 'response')
tab2 <- table(dresstrain$PLAS1, predict > 0.5)
tab2
```

```
##
##     FALSE TRUE
##   0   123    1
##   1    10    4
```

```
# Misclassification
incorrect <- 1 - sum(diag(tab2)) / sum(tab2)
correct<- 100*(1 - incorrect)
correct
```

```
## [1] 92.02899
```

```
#ROCR Curve
ROCRpred <- prediction(predict, dresstrain$PLAS1)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0
0.16
0.32
0.48
0.64
0.8